FISHER COLLEGE OF BUSINESS – HONORS THESIS

Using Twitter to Predict Investor Decisions

Jack Renner

INTRODUCTION

Since the beginning of the stock market in the 17th century, there has been much debate on how a stock is priced. Some say it comes from the stock's intrinsic value, or the cash flows, assets, dividends, etc. inherent of the company. Others suggest that investors are more irrational and that cognitive biases influence their decisions. In this research, I rely more on this second theory. How can one observe thoughts and biases on a macro basis though? The answer I use is Twitter. Past research has shown that sentiment expressed on Twitter proves a predictive relationship with the stock market in total. I want to expand this research by applying it to individual companies. By aggregating what people are Tweeting about and how they are Tweeting in regards to certain companies, I find that Twitter activity today is a statistically significant predictor of stock price tomorrow.

METHODS

1. Data Collection
   - An application called DiscoverText flagged Tweets based on keywords related to three companies: Starbucks, Home Depot, and Southwest Airlines.
   - I took a random sample of 50,000 Tweets from each set covering a span of 35, 24, and 46 days respectively.
   - From Yahoo! Finance I downloaded the relevant stock prices as well as the S&P 500 index prices.
   - Using WordStat, I was able to extract positive and negative words from each Tweet.
2. Variables
   - Dependent: Company open stock price.
   - Independent: Sum and mean of Tweets, words, positive words, negative words, etc. (total of 18 independent).
   - Controlled: S&P 500 Index open price.
3. Data Transformation
   - Stocks don't trade on the weekend, but people do Tweet.
   - Manually lagged the data by 1 day (Friday and Saturday Tweets not included in results).
4. Model Choice
   - Cross correlation to identify significant variables.
   - ARIMA model using those variables to create a detrended forecast.

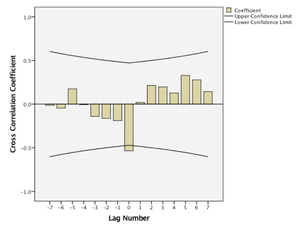
RESULTS

Hypothesis: The volume of Tweets about a company or the feelings people express toward a company over Twitter will show a relationship with the company's stock performance.

Step 1: Cross correlation between all independent variables and dependent variables to find related variables

Figure 1: Open Starbucks Price with Sum of Total Words

Here is an example of a significant cross correlation. It measures the relationship between the sum of total words on a day and the open price of Starbucks the following day. With the manual lag, a lag here of "0" is actually a lag of 1.



Step 2: Forecast stock values with significant Twitter variables while controlling for the S&P 500 index using an ARIMA model

Figure 2: Forecasted Starbucks Price Using Average Words Per Tweet Versus Observed Price

The average number of words in each Tweet is a significant negative predictor of Starbucks stock price, controlling for the S&P 500 index, with a significance value of 0.033. The ARIMA model parameters for auto-regression, integration, and moving average are [0, 1, 1], respectively

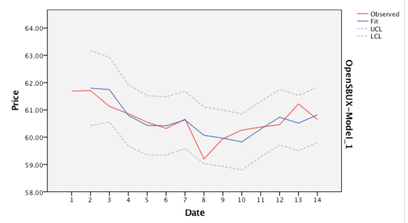


Figure 3: Forecasted Home Depot Price Using Total Number of Words Versus Observed Price

The total number of words in all the Tweets per day is a significant negative forecaster of Home Depot stock price, controlling for the S&P 500 index, with a significance value of 0.026. The ARIMA model parameters for auto-regression, integration, and moving average are [0, 0, 1], respectively

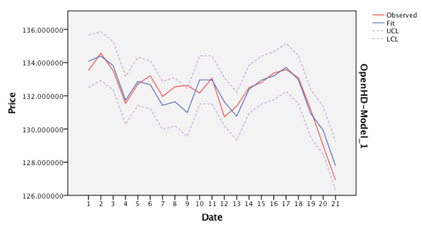
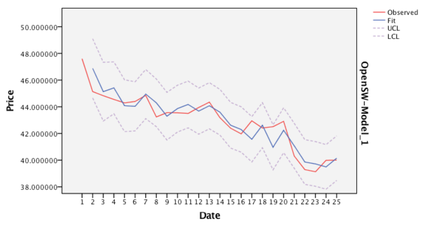


Figure 4: Forecasted Southwest Airlines Price Using Total Negative Words Versus Observed Price

The sum of negative words in all the Tweets per day is a significant positive forecaster of Southwest stock price, controlling for the S&P 500 index, with a significance value of 0.012. The ARIMA model parameters for auto-regression, integration, and moving average are [2, 1, 2], respectively



DISCUSSION

1. Starbucks
   - The more words people use in their Tweets on average indicates a stock drop the next day.
   - Possible interpretation: When people are feeling negatively about Starbucks, they will rant and say more in their Tweets.
2. Home Depot
   - The more total words Tweeted about Home Depot on a given day indicates a stock drop on the following day.
   - Possible interpretation: When people feel negatively about Home Depot, they tend to say a lot about it. Could be like the news – negative things tend to be talked about more.
3. Southwest Airlines
   - The more total negative words Tweeted about Southwest Airlines on a given day indicates a stock increase the following day.
   - Possible interpretation: Negativity surrounding the company results in overreaction in the stock market and so, the price the following day is showing a correction to the overreaction.

CONCLUSION

1. The implications of these results are that investors can use Twitter data to help forecast stock price in the future and potentially return a profit from it.
2. Limitations
   - Cannot prove why these particular Twitter factors are predictors
   - Cannot say why each forecast needed different factors
   - Cannot say if results would hold over longer time period or a different period
   - Losing two days of potentially valuable Twitter data per week
3. Opportunities for Future Research
   - Does this research apply to other companies?
   - Do companies with similar brands or products have similar forecast models?
   - Why can certain Twitter factors forecast stock price?

ACKNOWLEDGEMENTS

To my research advisor,

Dan McDonald,

Thank you for the countless hours you spent with me sifting through all this data. Your knowledge, patience, and genuine joy learning new things are inspirational to me. I couldn't have done it without you!